Research Article | Vol. 12, No. 5 / 1 May 2021 / Biomedical Optics Express 2684
Biomedical Optics EXPRESS

Automatic detection of retinopathy with optical coherence tomography images via a semi-supervised deep learning method

YUEMEI LUO,[1] QING XU,[2] RUIBING JIN,[2] MIN WU,[2,*] AND LINBO LIU[1,3]

[1] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798, Singapore
[2] Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore, 138632, Singapore
[3] School of Chemical and Biomedical Engineering, Nanyang Technological University, Singapore, 637459, Singapore
[*] wumin@i2r.a-star.edu.sg

Abstract: Automatic detection of retinopathy via computer vision techniques is of great importance for clinical applications. However, traditional deep learning based methods in computer vision require a large amount of labeled data, which are expensive and may not be available in clinical applications. To mitigate this issue, in this paper, we propose a semi-supervised deep learning method built upon pre-trained VGG-16 and virtual adversarial training (VAT) for the detection of retinopathy with optical coherence tomography (OCT) images. It only requires very few labeled and a number of unlabeled OCT images for model training. In experiments, we have evaluated the proposed method on two popular datasets. With only 80 labeled OCT images, the proposed method can achieve classification accuracies of 0.942 and 0.936, sensitivities of 0.942 and 0.936, specificities of 0.971 and 0.979, and AUCs (Area under the ROC Curves) of 0.997 and 0.993 on the two datasets, respectively. When comparing with human experts, it achieves expert level with 80 labeled OCT images and outperforms four out of six experts with 200 labeled OCT images. Furthermore, we also adopt the Gradient Class Activation Map (Grad-CAM) method to visualize the key regions that the proposed method focuses on when making predictions. It shows that the proposed method can accurately recognize the key patterns of the input OCT images when predicting retinopathy.

1. Introduction

Optical coherence tomography (OCT) has been widely adopted by ophthalmologists for the detection of retinopathies due to its high resolution, contactless and nondestructive testing properties [1,2]. Based on the textural and morphological variations, it is able to detect major retinopathies, like age-related macular degeneration (AMD) and diabetic macular edema (DME), which generally lead to vision loss. The AMD can be further divided into "dry" AMD and "wet" AMD (also known as choroidal neovascularization (CNV)), where the typical feature of DRUSEN can be seen. With huge number of scanned OCT images, i.e., around 30 millions per year [3], manual identification of OCT images is a huge burden for ophthalmologist. The identification task becomes even more challenging in developing countries where the number of qualified ophthalmologists is inadequate.

Recently, deep learning has achieved great success in many challenging areas, such as image classification and nature language understanding [4]. One of the most popular deep learning algorithms is the convolutional neural network (CNN). CNN is designed for image processing and thus it has been widely used for the recognition of OCT images [5–8]. Lee et al. designed a

#418364 https://doi.org/10.1364/BOE.418364
Journal © 2021 Received 22 Dec 2020; revised 27 Mar 2021; accepted 2 Apr 2021; published 13 Apr 2021

Biomedical Optics EXPRESS

CNN-based method, i.e., VGG-16, to classify normal verses AMD based on OCT images [9]. In their experiments, more than eighty thousand OCT images were unitized to train the model. In [10], the authors applied multiple CNN-based deep learning methods, such as ResNet, DenseNet, CliqueNet, etc., for OCT image classification. Two public datasets were utilized for evaluation, where 80% of data are employed for model training. Rasti et al. proposed a multi-scale CNN ensemble method for macular OCT classification [11]. They achieved the precision over 98% and the AUC (Area under the ROC Curve) over 0.99 in five-fold cross-validation. Li et al. proposed an ensemble of ResNet for retinopathy detection with OCT images, where 13,192 OCT images were employed for model training [12]. Deep learning-based retinopathy detection with OCT images has already achieved remarkable performance, even comparable to ophthalmologists [13,14]. However, the success of deep learning algorithms heavily relies on large amount of labeled data for model training, which is expensive and requires lots of human effort.

Pretraining on other image datasets and fine-tuning is a common way to alleviate the requirement of huge data [15]. Karri et al. proposed a pretrained method to classify OCT images with DME and AMD [15]. Specifically, a pretrained CNN, i.e., GoogLeNet, was fine-tuned on OCT images for classification. A similar idea can be found in [13,16], where a pretrained VGG-16 network was adopted and fine-tuned for retinopathy detection with OCT images. Even with pretrained networks, several thousands of OCT images are still required to fine-tune the networks for a satisfactory performance. This limits the applicability of deep learning-based automatic retinopathy detection in real life.

To address these critical issues, semi-supervised deep learning can be a good candidate. It intends to use both labelled and unlabelled data for model training. Ding et al. proposed a pseudo-labeling method for semi-supervised deep learning, which attempts to annotate unlabelled data with pseudo labels and then train the model with both the labelled data and unlabelled data with pseudo labels [17]. Yao et al. developed a temporal ensembling based semi-supervised deep learning method [18] with two losses, i.e., supervised loss for labeled data and consistency loss for unlabelled data. In particular, the consistency loss calculates the difference between the current prediction and the mean of previous predictions for unlabelled data. Another popular semi-supervised deep learning method is mean teacher [19]. It designs a teacher model based on the moving average of the weights of the original model (also known as student). And then, a consistency loss is calculated based on the difference between the teacher and the student outputs for the unlabelled data. These models have some limitations: 1) the quality of pseudo labels greatly influences the model performance; 2) the consistency loss may be unstable, especially at the early training stage, resulting limited performance.

In this paper, we propose a semi-supervised deep learning method built upon virtual adversarial training (VAT) [20] for automatic detection of retinopathy with OCT images. Since unlabeled OCT images can be easily obtained, we attempt to use large amount of unlabeled OCT images to assist the training of deep leaning algorithms. Specifically, we propose a pretrained VGG-16 with VAT algorithm to leverage on both labeled and unlabeled OCT images for retinopathy detection. The cross-entropy loss calculated with the labeled images and the VAT loss calculated with the unlabeled images are combined to update the parameters of the deep learning algorithm with gradient-based methods. Besides, we also adopt the Gradient Class Activation Map (Grad-CAM) [21] to visualize the regions of the input images that the model focus on when making predictions. As such, it is able to check whether the proposed method can recognize the key patterns during prediction. To evaluate the performance of the proposed method, we use two widely used public datasets [22,23]. With only 80 labeled OCT images, the proposed method achieves classification accuracies of 0.942 and 0.936 on the two datasets, which attains the human expert level. By using Grad-CAM, we show that the proposed method can find the key patterns when predicting retinopathy.

Biomedical Optics EXPRESS

2. Methodology

2.1. Overview

The proposed framework for retinopathy detection is shown in Fig. 1. Firstly, the VGG-16 model is pretrained on the popular ImageNet dataset [24]. Note that the standard VGG-16 network was designed for the classification task with 1,000 classes. Therefore, the last layer, i.e., the softmax layer, contains 1,000 outputs, corresponding to the predicted probabilities for 1,000 classes. In our case, the number of outputs for the softmax layer needs to be changed to the category number of retinopathies.

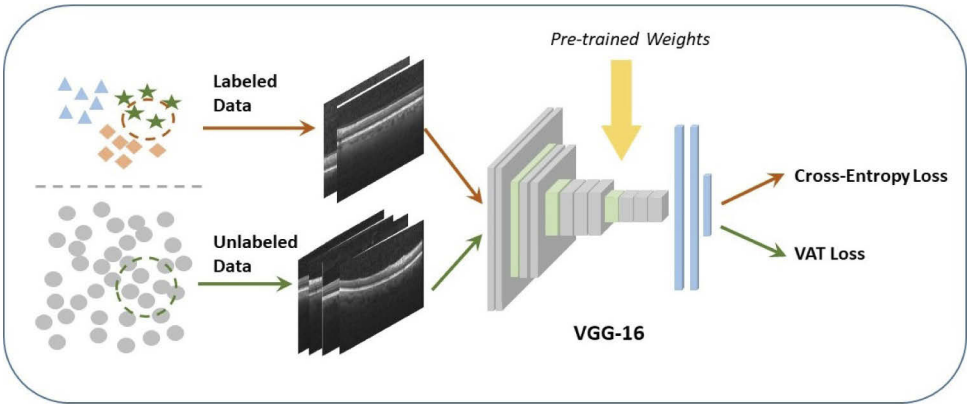


Fig. 1. Proposed deep semi-supervised learning framework for retinopathy detection with OCT images.

After the model pretraining, the next step is to fine-tune the parameters of the model with both labeled and unlabeled data. Specifically, we replace the last layer of pretrained VGG-16 (i.e., the softmax layer with 1,000 outputs) by a softmax layer with C outputs where C refers to the types of diseases, e.g., normal, AMD, DME, etc. The weights of this new softmax layer is randomly initialized. Next, the model is fine-tuned using the small amount of labelled OCT images with large amount of unlabelled ones. Note that, the pretraining provides a good initialization for fast and better convergence of the model. For the labeled and unlabeled data, a cross-entropy loss and a VAT loss are calculated, respectively. Then, both losses are combined to obtain the final loss of the model. Finally, the stochastic gradient decent (SGD) calculated upon the final loss is adopted to optimize the parameters of the model for retinopathy detection. The detailed introduction of the VGG-16 with pretraining and the VAT will be presented in the following subsections.

2.2. Pretrained VGG-16 for OCT images

VGG-16, a very deep CNN model, has been widely used for image classification tasks [25]. Figure 2 shows the detailed structure of the VGG-16 for the detection of retinopathy with OCT images. Firstly, the original OCT B-scans are resized to the dimension of $224 \times 224 \times 3$ which is the input size of VGG-16. The resized OCT B-scans will go through several convolutional layers with a filter size of 3×3 and a stride size of 1, followed by max-pooling layers with a pooling size of 2×2 with a stride size of 2. The number of filters for each convolutional layer can be found in Fig. 2. Then, a flatten layer is utilized to convert the data into a one-dimensional feature vector after the last max-pooling layer. The one-dimensional feature vector will be fed into two fully-connected (FC) layers with a hidden size of 4096 and a softmax classification layer with C outputs. The activation functions of the convolutional layers and fully-connected layers are the popular rectified linear unit (ReLU).

Biomedical Optics EXPRESS

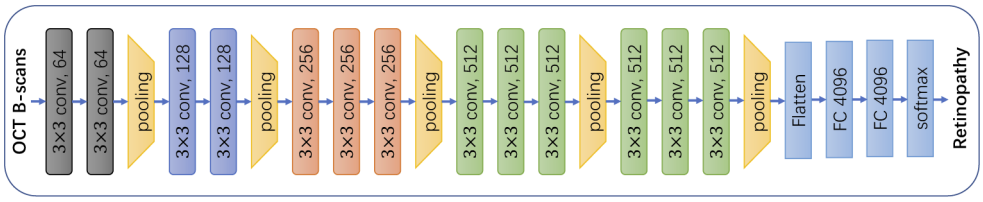


Fig. 2. The VGG-16 for retinopathy detection with OCT B-scans.

To train a very deep CNN model from scratch (i.e., VGG-16), it generally requires large amount of labeled data. Here, we adopt the idea of pretraining which intends to train the model on a large public dataset with labels. Specifically, we train the VGG-16 on the ImageNet dataset. The pretrained VGG-16 thus has a strong capability of feature learning on images. When applying this pretrained VGG-16 for the detection of retinopathy on OCT images, it requires less labeled images and converges faster in model training.

2.3. *Virtual adversarial training*

Due to the high cost of data annotation and the lack of ophthalmologists in many countries, especially the developing countries, collecting a large labeled OCT image dataset to train a deep learning model for automatic detection of retinopathies is not realistic. However, unlabeled OCT images are widely available, e.g., around 30 millions of OCT images have been generated per year [26]. In this work, we attempt to develop a deep learning algorithm (built upon the pretrained VGG-16) with very few labeled and large amount of unlabeled OCT images.

We adopt the idea of local distributional smoothness (LDS) which indicates the smoothness of the output distributions of a model with respect to the inputs [20]. In another words, the changes of the model outputs should be small with small input perturbations, resulting a robust model with good generalization performance. Virtual adversarial training (VAT) is an effective way to achieve LDS of a model [20]. It intends to find a small perturbation of an input, such that the output distributions of the model have the largest change. Then, the model is trained to minimize this change due to the perturbed input. Specifically, given an input data $\mathbf{x}$ (e.g., an OCT image), a new sample, $\tilde{\mathbf{x}}$ will be generated by adding a small perturbation μ to the original input $\mathbf{x}$, i.e., $\tilde{\mathbf{x}} = \mathbf{x} + \mu$. Among all the possible perturbations μ, the adversarial perturbation, denoted as μ_{adv}, will lead to the largest change over the model outputs. Here, the change of the model outputs is measured by using the KL divergence. The μ_{adv} can be obtained by solving the following equation:

$$\begin{aligned} \Delta_{\mathrm{KL}}(\mu, \mathbf{x}, \theta) &= \mathrm{KL}\left[p(y|\mathbf{x}, \theta)||p(y|\mathbf{x} + \mu, \theta)\right], \\ \mu_{adv} &= \underset{\mu}{\operatorname{argmax}} \left\{\Delta_{\mathrm{KL}}(\mu, \mathbf{x}, \theta); \|\mu\|_2 \leq \epsilon\right\}, \end{aligned} \tag{1}$$

where θ is the model parameters, y is the model output, and $\|\mu\|_2 \leq \epsilon$ is to ensure that the perturbation μ is smaller than ϵ which is pre-defined hyper-parameter. After obtaining the adversarial perturbation μ_{adv}, the objective of VAT is to minimize the KL divergence of the model outputs, resulting a robust model with LDS towards various perturbations. Thus, the VAT loss can be defined as follows:

$$\mathrm{VAT}_{loss} = \Delta_{\mathrm{KL}}(\mu_{adv}, \mathbf{x}, \theta). \tag{2}$$

Based on the above analysis, it is clear that the calculation of the VAT loss only requires the model outputs, but no labels. It is different from conventional adversarial training which requires the true labels. Therefore, the VAT loss can be used for semi-supervised learning where large number of unlabeled data are available.

Biomedical Optics EXPRESS

In this work, we assume that very few labeled and large number of unlabeled OCT images are available for retinopathy detection. For the labeled OCT images, the cross-entropy (CE) loss is calculated during model training, which can be expressed as

$$\mathrm{CE}_{loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\log\hat{y}_{ic}, \tag{3}$$

where N is the total number of labeled OCT images, C is the number of retinopathies, y_{ic} represents the true probability of the sample i and the class c, and $\hat{y}_{ic}$ represents the predicted probability. The VAT loss in Eq. (2) is then employed for unlabeled OCT images. Hence, the overall loss of the proposed method can be expressed as

$$\mathrm{CE}_{loss}(\mathbf{x}_l) + \alpha\mathrm{VAT}_{loss}(\mathbf{x}_{ul}), \tag{4}$$

where $\mathbf{x}_l$ and $\mathbf{x}_{ul}$ denote the labeled and unlabeled OCT images, and α is the hyper-parameter to control the contribution of the CE loss and the VAT loss.

2.4. Grad-CAM for visualization

For the automatic detection of retinopathy via deep learning, a major concern is how to interpret the predictions of the model, so that the ophthalmologists will be confident on the predictions. Here, we adopt a gradient-based method, i.e., Grad-CAM [21], which is able to visualize the regions of inputs (also known as saliency map) that are "vital" for the current prediction. As the retinopathy has clear patterns on OCT images, we can verify the model by checking whether it can focus on these patterns when predicting different retinopathies.

The basic idea of Grad-CAM for the visualization of the proposed method is to use the gradient information flowing into the last convolutional layer of the VGG-16 to understand each neuron when making predictions. Assume that y_c is the score (before the softmax layer) of the class of interest, and A_k^{ij} is the pixel at i-th row and j-th column of the k-th feature map of the last convolutional layer, we can calculate the weight of each feature map based on the gradient information, shown as

$$w_k = \frac{1}{Z}\sum_i\sum_j \frac{\partial y_c}{\partial \mathbf{A}_k^{ij}}, \tag{5}$$

where Z is the total number of pixels for the k-th feature map. After obtaining the weights for all the feature maps, the saliency map can be expressed as

$$S_{\text{Grad-CAM}} = \mathrm{ReLU}\left(\sum_k w_k\mathbf{A}^k\right). \tag{6}$$

Here, the ReLU helps us to focus on the pixels that have positive relationship with the class of interest. With the saliency map, we are able to understand which regions of an OCT image play a key role when predicting different retinopathies. This can be further verified by ophthalmologists, such that users are confident when using the proposed method for automatic detection of retinopathy with OCT images.

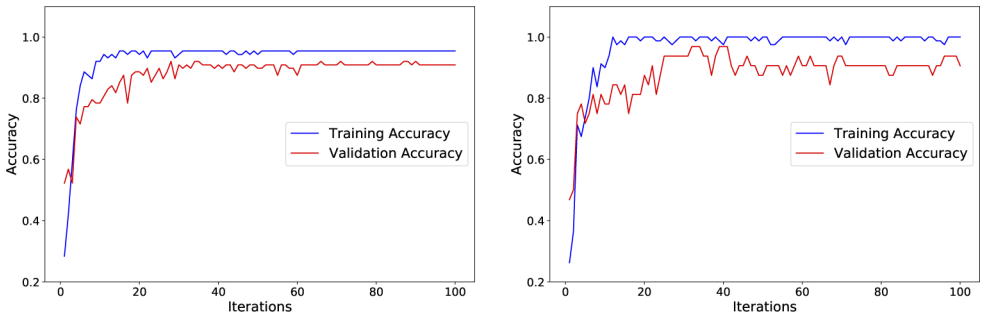
2.5. Implementation details

For the proposed method, stochastic gradient descent (SGD) with a learning rate of 0.001 and a momentum of 0.9 is adopted for parameter optimization in training. Here, we use batch training. The batch size is 8 for the labeled data and the batch sizes are 16 and 32 for the unlabelled data in the two datasets (namely BOE and CELL respectively). Let's use the training procedure on the BOE dataset as as example. In particular, for each individual step, given a batch of labelled OCT

Biomedical Optics EXPRESS

images (8 images) and a batch of unlabelled OCT images (16 images), the CE loss and the VAT loss can be calculated. Then, the overall loss can be obtained based on Equation (4), which is back-propagated for parameter optimization with SGD. Since there are much more unlabelled OCT images than labelled ones, when looping over all the labelled OCT images (also known as one training epoch), the unlabelled data still have many batches left. Then, for the next training epoch, the labelled data starts at the first batch, while the unlabelled data starts at the next batch for the remaining batches, until last batch. After looping over all the batches in the unlabelled data, it goes to the first batch for the next training step. Note that, for the CELL dataset which has too many unlabelled data, we may not be able to loop over all the unlabelled OCT images even when we have already trained on the labelled data for many epochs.

In this paper, ϵ is set to 3 in Equation (1) and α is set to 2 in Equation (4). These hyper-parameters are determined based on the grid search on the validation set. The training is stopped when the performance on the validation set starts to degrade. The detailed training procedure is shown in Fig. 3. It can be found that the proposed method can converge very fast as the pretraining provides a good initialization. Moreover, the techniques of dropout and batch-normalization have been adopted to prevent overfitting.



(a) BOE dataset (b) CELL dataset

Fig. 3. Training and validation accuracies in the training procedure for the two datasets.

The code is written under the PyTorch platform with Python 3 and running on a NVIDIA GeForce RTX 2080 Ti GPU. We have released the code of the proposed method in GitHub with the following link: https://github.com/xuqing88/Pytorch-SSDL-OCT.

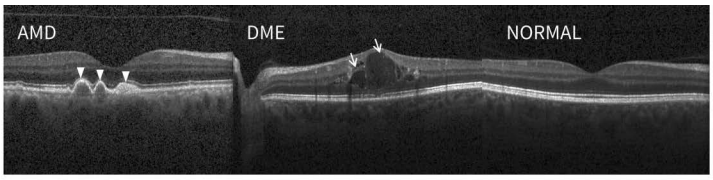
3. Evaluation

3.1. Data description

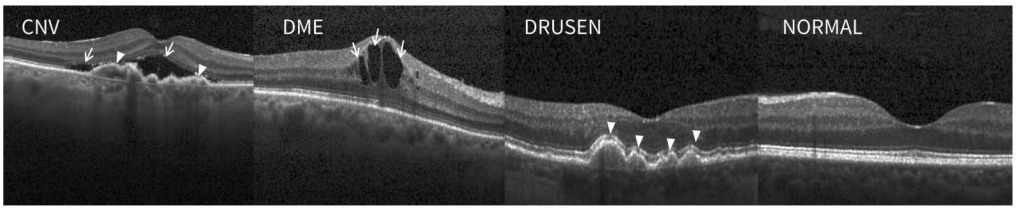
In this paper, we use two popular retinopathy detection datasets to evaluate the performance of the proposed method. The first dataset (denoted as BOE dataset) was collected by Srinivasan et al. [22] from 45 subjects in three Universities, i.e., Duke University, Harvard University and the University of Michigan. It contains 723 images with AMD, 1,101 images with DME, and 1,407 NORMAL images, which were collected from 15 normal patients, 15 patients with AMD, and 15 patients with DME. Note that, each patient has one volume scan with different number of B-scans. For the BOE dataset, the number of outputs in the softmax layer, C, equals to 3 and it is a three-class classification task. The second dataset (denoted as CELL dataset) was obtained from [23]. Totally, 84,484 OCT images were collected from 5,319 subjects, where the number of CNV, DME, DRUSEN and NORMAL images are 37,455, 11,498, 8,866, and 26,565, respectively. For the CELL dataset, C equals to 4 and it is a four-class classification task. Fig. 4 shows some typical OCT images with different retinopathies in the two datasets. Since the input size of VGG-16 is $224 \times 224 \times 3$, the first step is to resize all the OCT images into the dimension

Biomedical Optics EXPRESS

of $224 \times 224 \times 3$. Note that, the original size of the OCT images is varying. This is the only preprocessing that needs to be done for the proposed method.



(a) The BOE dataset with AMD, DME and NORMAL



(b) The CELL dataset with CNV, DME, DRUSEN and NORMAL

Fig. 4. An illustration of OCT images with different retinopathies in the two datasets. The arrows indicate the critical parts, such as neovascular membrane and retinal fluid, for the recognition of retinopathy.

In this paper, we consider a very challenging and practical scenario where only very few labeled and large amount of unlabeled OCT images are used for training the proposed method. Specifically, we randomly select 80 labeled OCT images, which account for around 0.25% and 0.095% of total OCT images in BOE and CELL datasets respectively, for training the proposed method. Note that the training also includes 1,357 unlabeled OCT images for the BOE dataset and 83,324 unlabeled OCT images for the CELL dataset. Another 80 randomly selected OCT images are utilized for validation. For testing, we use 924 and 1,000 OCT images in BOE and CELL datasets, respectively. Note that we perform patient level random selection to guarantee that the scans from the same patient appear in one set only (i.e., training, validation or test set). Specifically, training and validation data are from 33 and 4,686 subjects for the BOE and CELL datasets, respectively. While the test data are from another 12 and 633 subjects for the BOE and CELL datasets, respectively.

3.2. Experimental setup

To evaluate the performance of the proposed method, we have compared it with some benchmark approaches in the literature, including the SVM with Histogram of Oriented Gradients (HOG) features [22], normal CNN [27], AlexNet [28] with and without pretraining, ResNet-18 [10] with and without Pretraining, VGG-16 [16] with and without pretraining. Note that these supervised learning methods can only use the labeled OCT images for model training, i.e., 80 labeled OCT images. We also conducted a comparison with some advanced semi-supervised deep learning methods, i.e., pseudo-labeling [17], temporal ensembling [18], and mean teacher [19]. The pretrained VGG-16 is adopted to combine with these semi-supervised techniques for fair comparison.

In order to quantify the performance of different methods, we adopt the evaluation metrics of classification accuracy, sensitivity (also known as recall), specificity, AUC and ROC curve, which are widely used in the literature [10,16,29].

Biomedical Optics EXPRESS

The definition of overall accuracy can be straightforward, which can be expressed as

$$Accuracy_{overall} = \frac{Number\ of\ correctly\ classified\ samples}{Total\ number\ of\ samples} = \frac{TP_A + TP_B + TP_C}{N} \tag{7}$$

where N is the total number of samples, and TP_A, TP_B, and TP_C are the numbers of correctly classified samples (True Positive) for the classes of A, B and C respectively.

Nevertheless, the sensitivity, specificity, AUC and ROC curve are originally defined in binary classification. In order to use these evaluation metrics for multi-class classification, we adopt the general strategy of *one-vs-rest* [30].

To explain the calculation of sensitivity and specificity, let's use a toy example for three-class classification (namely class A, B and C). We can define three *one-vs-rest* scenarios, i.e., *A-vs-rest*, *B-vs-rest*, and *C-vs-rest*. Figure 5 shows the calculations of sensitivity and specificity for the scenarios of *A-vs-rest*, denoted as $Sensitivity_A$ and $Specificity_A$, *B-vs-rest*, denoted as $Sensitivity_B$ and $Specificity_B$, and *C-vs-rest*, denoted as $Sensitivity_C$ and $Specificity_C$. Here, we use "micro-average" strategy for the calculation of overall sensitivity and specificity. In particular, the reported sensitivity and specificity are weighted averages over all the *one-vs-rest* scenarios. The detailed calculations are shown in Eq. (8) and (9), where N_A, N_B, and N_C represent the total numbers of samples in the classes of A, B and C respectively, and $N = N_A + N_B + N_C$.

$$Sensitivity_{micro} = \frac{N_A}{N} * Sensitivity_A + \frac{N_B}{N} * Sensitivity_B + \frac{N_C}{N} * Sensitivity_C \tag{8}$$

$$Specificity_{micro} = \frac{N_A}{N} * Specificity_A + \frac{N_B}{N} * Specificity_B + \frac{N_C}{N} * Specificity_C \tag{9}$$

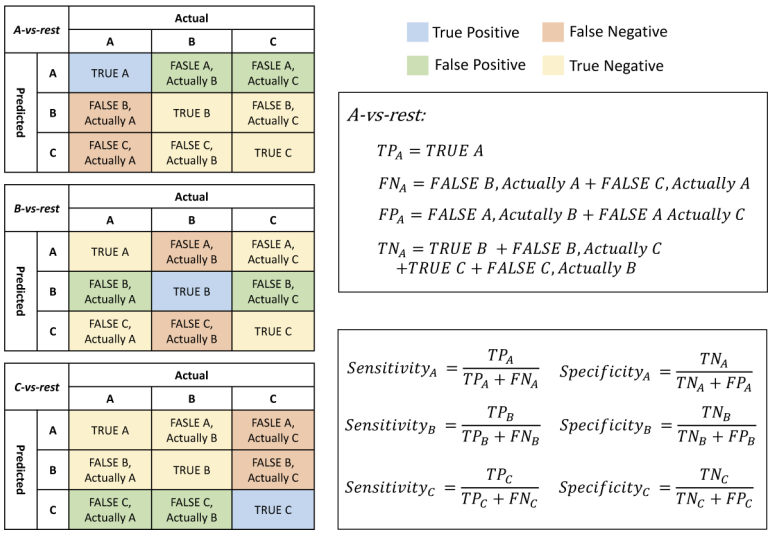


Fig. 5. The calculation of sensitivity and specificity for a toy example of three-class classification.

We also adopt the *one-vs-rest* strategy to draw the ROC curve for each method. Specifically, with the *one-vs-rest* strategy, it will generate C ROC curves (C equals to the number of classes) for each method. Then, the final ROC curve can be obtained by averaging over all the C ROC curves. The corresponding AUC is the area under the final ROC curve.

Biomedical Optics EXPRESS

3.3. Comparison with benchmark approaches

The experimental results are shown in Table 1 and Table 2. It can be found that the VGG-16 with pretraining has a superior performance over all the other benchmark approaches on both datasets. This indicates the effectiveness of the VGG-16 network and the pretraining scheme for the detection of retinopathy with OCT images. The semi-supervised learning algorithms can effectively improve the performance of retinopathy detection. Moreover, the proposed semi-supervised deep learning method significantly outperforms all the benchmark approaches with accuracies of 0.942 and 0.936, sensitivities of 0.942 and 0.936, specificities of 0.971 and 0.979, and AUCs of 0.997 and 0.993, on the two datasets. Note that, we only use 80 labeled OCT images for training the proposed method. The standard deviation of the proposed method is also small, indicating its robustness. Since the SVM does not contain random components, the standard deviation of SVM is not available.

Table 1. The experimental results on the BOE dataset. ACC refers to the classification accuracy. STD refers to the standard deviation of multiple runs. The p-value of 1.00E-07 refers to 1.0×10^{-7}.

Method	ACC	STD	p-value	Sensitivity	Specificity	AUC	Train Time (s)	Test Time (s)
SVM+HOG	0.570	-	1.00E-07	0.570	0.785	0.774	103	1.38
CNN	0.802	0.075	4.00E-03	0.802	0.901	0.950	12,591	7.38
ResNet-18	0.601	0.051	9.16E-06	0.601	0.800	0.803	8,213	5.47
ResNet-18+Pretrain	0.840	0.025	2.59E-03	0.840	0.920	0.957	758	5.44
AlexNet	0.523	0.035	1.00E-07	0.523	0.761	0.618	7,750	5.47
AlexNet+Pretrain	0.754	0.019	7.49E-06	0.754	0.877	0.841	707	5.48
VGG-16	0.629	0.066	3.36E-06	0.629	0.814	0.791	17,329	8.32
VGG-16+Pretrain	0.855	0.028	5.83E-03	0.855	0.927	0.974	1,577	8.31
Pseudo-Labeling	0.906	0.027	0.0039	0.906	0.951	0.984	5,087	8.25
Temporal Ensembling	0.893	0.054	0.0120	0.893	0.947	0.957	6,282	8.19
Mean Teacher	0.918	0.043	0.0110	0.918	0.963	0.988	7,553	8.45
Proposed	0.942	0.027	-	0.942	0.971	0.997	6,145	8.58

To demonstrate the significance of the proposed method over benchmark approaches, p-values of the proposed method over other methods are also shown in Table 1 and Table 2. It can be observed that p-values are all less than 0.05, which indicates the significance of the proposed method. We also present the training and testing time for all the approaches. Even though most of methods requires a long time (up to 5 hours) for model training, this tedious training process only needs to be done once, which is still acceptable for real applications. The testing time (for all the testing samples, i.e., 924 for BOE and 1,000 for CELL) of all the approaches is quite small. The inference time for a single OCT image is less than 10 ms for the proposed method, which is adequate for real applications.

We also implemented the supervised baseline methods with all labeled images in the training set (i.e., 1,437 labelled OCT images for BOE and 83,404 labelled OCT images for CELL), which can be treated as a higher bound. The results are shown in Table 3 and 4. For the BOE dataset, the proposed method (accuracy 0.942) can outperform most of the supervised baseline methods with all the labeled OCT images for training. It performs slightly worse than the VGG-16 with pretraining. For the CELL dataset, all the supervised baseline methods (except SVM) significantly outperform the proposed method (accuracy 0.936). It is worth noting that for the BOE dataset, the supervised baseline methods use 1,437 labeled OCT images for training, which is around 18 times more labeled images than our semi-supervised learning method (with only 80

Biomedical Optics EXPRESS

Table 2. The experimental results on the CELL dataset. ACC refers to the classification accuracy. STD refers to the standard deviation of multiple runs. The p-value of 1.00E-07 refers to 1.0×10^{-7}.

Method	ACC	STD	p-value	Sensitivity	Specificity	AUC	Train Time (s)	Test Time (s)
SVM+HOG	0.522	-	1.00E-07	0.522	0.841	0.783	106	1.66
CNN	0.556	0.040	1.00E-07	0.556	0.852	0.815	13,314	6.25
ResNet-18	0.569	0.038	1.00E-07	0.569	0.856	0.814	7,688	4.19
ResNet-18+Pretrain	0.826	0.036	7.80E-05	0.826	0.942	0.965	722	4.20
AlexNet	0.432	0.026	1.00E-07	0.432	0.811	0.685	6,976	4.02
AlexNet+Pretrain	0.837	0.030	1.10E-05	0.837	0.946	0.971	655	4.03
VGG-16	0.527	0.060	2.00E-07	0.527	0.842	0.797	18,821	7.18
VGG-16+Pretrain	0.844	0.025	2.50E-05	0.844	0.948	0.976	1,723	7.16
Pseudo-Labeling	0.916	0.018	0.0063	0.916	0.972	0.991	3,767	6.52
Temporal Ensembling	0.863	0.024	0.0004	0.863	0.955	0.977	4,950	7.68
Mean Teacher	0.881	0.016	0.0067	0.881	0.961	0.980	5,832	7.14
Proposed	0.936	0.007	-	0.936	0.979	0.993	4,505	7.08

labeled OCT images). While, for the CELL dataset, the supervised baseline methods use 83,404 labeled OCT images for training, which is around 1,000 times more labeled images than our semi-supervised learning method. That is why all the supervised learning methods including these without pretraining can achieve very good classification performance and significantly outperform the proposed method. But in real life, it is almost impossible to obtain such a huge labelled dataset for model training. Even labeling thousands of OCT images can be very challenging for resource-limited regions. That is why the proposed method can be of great significance in real applications.

Table 3. The experimental results of supervised baseline methods on the BOE dataset with all labeled training samples.

Method	ACC	Sensitivity	Specificity	AUC
SVM+HOG	0.742	0.742	0.871	0.806
CNN	0.928	0.928	0.964	0.993
Resnet-18	0.765	0.765	0.881	0.911
Resnet-18+Pretrain	0.908	0.908	0.954	0.981
AlexNet	0.582	0.582	0.791	0.774
AlexNet+Pretrain	0.801	0.801	0.900	0.866
VGG-16	0.637	0.637	0.819	0.782
VGG-16+Pretrain	0.946	0.946	0.973	0.994

The ROC curves of the all the approaches on the two datasets are demonstrated in Fig. 6. It is clear that the proposed semi-supervised deep learning method performs the best, which is consistent with other evaluation metrics in Tables 1 and 2. The confusion matrices of the proposed method are illustrated in Fig. 7. It can be found that the NORMAL class is relatively easy to be detected due to the clean patterns (see Fig. 4). Considering that the DME has more complicated patterns, the recognition of DME is relatively challenging in both datasets.

Biomedical Optics EXPRESS

Table 4. The experimental results of supervised baseline methods on the CELL dataset with all labeled training samples.

Method	ACC	Sensitivity	Specificity	AUC
SVM+HOG	0.798	0.798	0.933	0.947
CNN	0.972	0.972	0.991	0.999
Resnet-18	0.978	0.978	0.993	1.000
Resnet-18+Pretrain	0.994	0.994	0.998	1.000
AlexNet	0.978	0.978	0.993	1.000
AlexNet+Pretrain	0.991	0.991	0.997	1.000
VGG-16	0.984	0.984	0.995	1.000
VGG-16+Pretrain	0.998	0.998	0.999	1.000

3.4. Impact of the number of labeled OCT images

Here, we investigate the performance of the VGG-16 with pretraining and the proposed semi-supervised deep learning method with different numbers of labeled OCT images. Figure 8 shows the experimental results. Specifically, we test the two methods with 40, 60, 80, 100, 120, 160 and 200 labeled OCT images. It is obvious that the performance of the models improves with more labeled OCT images. If the number of OCT images is too few, e.g., 40, the models may not be able to converge during training, leading to a poor performance. When the number of labeled OCT images is larger than 60, the accuracies of the proposed model on both datasets are higher than 0.9, which is very crucial in real application where the number of labeled images is very limited. In all scenarios, the proposed method outperforms the VGG-16 with pretraining, which indicates the effectiveness of the proposed method.

3.5. Impacts of the hyper-parameters α and ϵ

For our proposed method, α in Equation (4) is a key hyper-parameter, which controls the contributions of the CE loss and the VAT loss. Specifically, we tested α with values of [0, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0, 20.0] on the two datasets. We also investigated the impact of hyper-parameter ϵ, which controls the magnitude of the perturbation in Equation (1), on model performance.

The results are shown in Fig. 9. Note that $\alpha = 0$ means that the VAT loss is not used, which turns out to be the VGG-16 with pretraining. It can be found that the performance of the proposed method improves with larger α values (more contributions of the VAT loss) at the beginning. This indicates the effectiveness of the VAT loss on unlabeled data for enhancing the performance of retinopathy detection. But when the α value is too large, the performance of the proposed method degrades. This is because the optimization will be mainly based on the VAT loss on the unlabeled data, which lacks the correct supervision from the labeled data. Therefore, a careful selection of this hyper-parameter is of great importance for the proposed method. In this paper, we have chosen $\alpha = 2$ for both datasets. According to Fig. 9, too small or large perturbation will also lead to a degraded performance of the proposed method. In this paper, we have chosen $\epsilon = 3$ for both datasets.

3.6. Comparison with human experts

For the CELL dataset, Kermany et al. also asked six human experts with significant experience in an academic ophthalmology center to classify 1,000 OCT images [13]. Here, we attempt to compare the proposed method with these human experts for identifying different retinopathies with OCT images. The results are presented in Fig. 10, where the proposed method with different number of labeled OCT images, i.e., 80, 160 and 200, are selected for comparison. It can be

Biomedical Optics EXPRESS

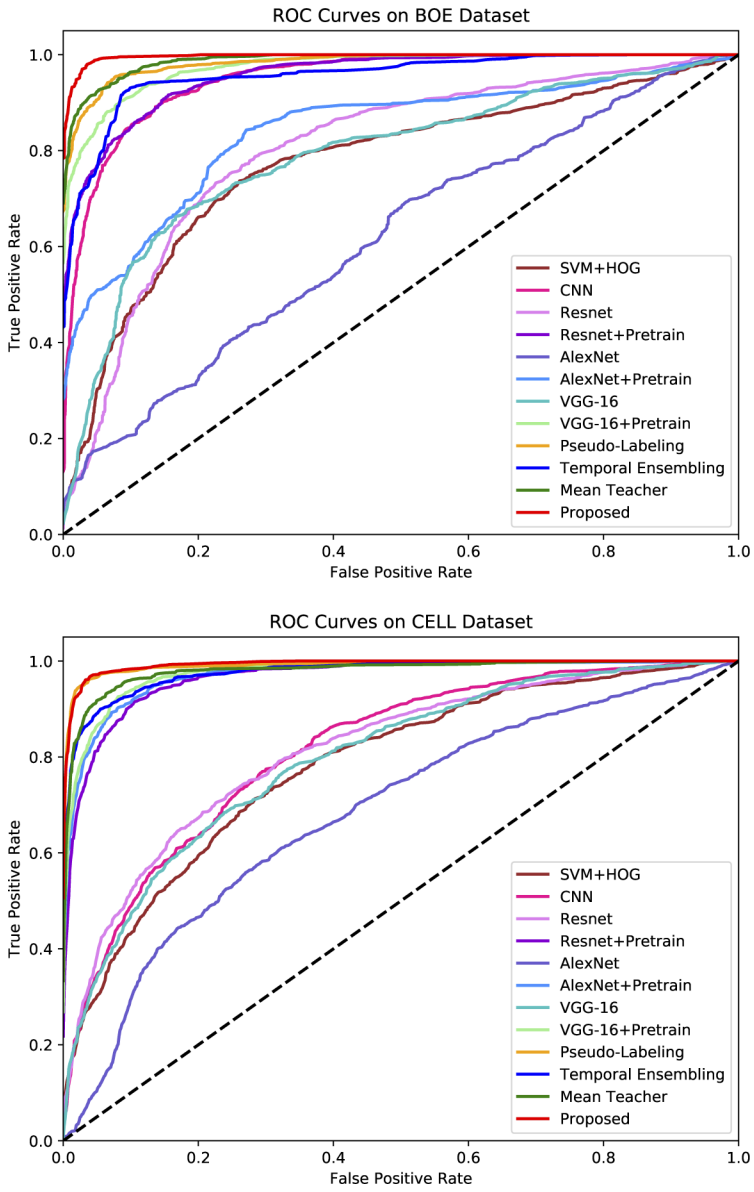


Fig. 6. The ROC curves of all the approaches on the two datasets.

Biomedical Optics EXPRESS

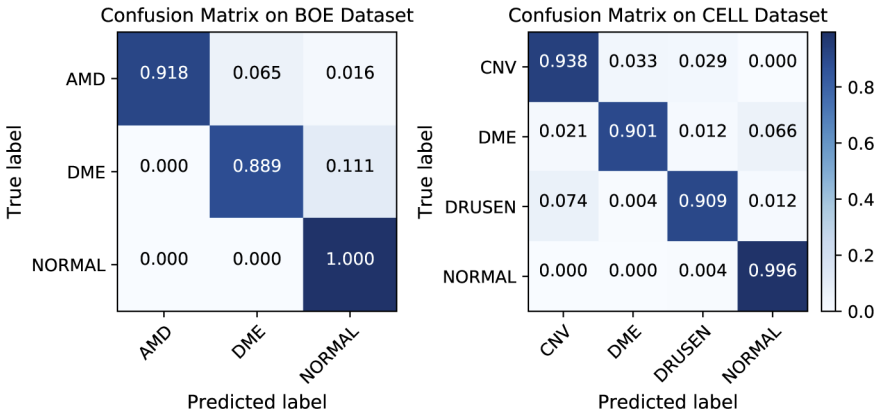


Fig. 7. Confusion matrices of the proposed method on the two datasets. The values of the diagonal elements represent the classification accuracies for each class. And the rest refer to the misclassification rates for each scenario. For example, the value at the first row and second column in the left figure represents that 6.5% of samples with AMD have been wrongly classified as DME.

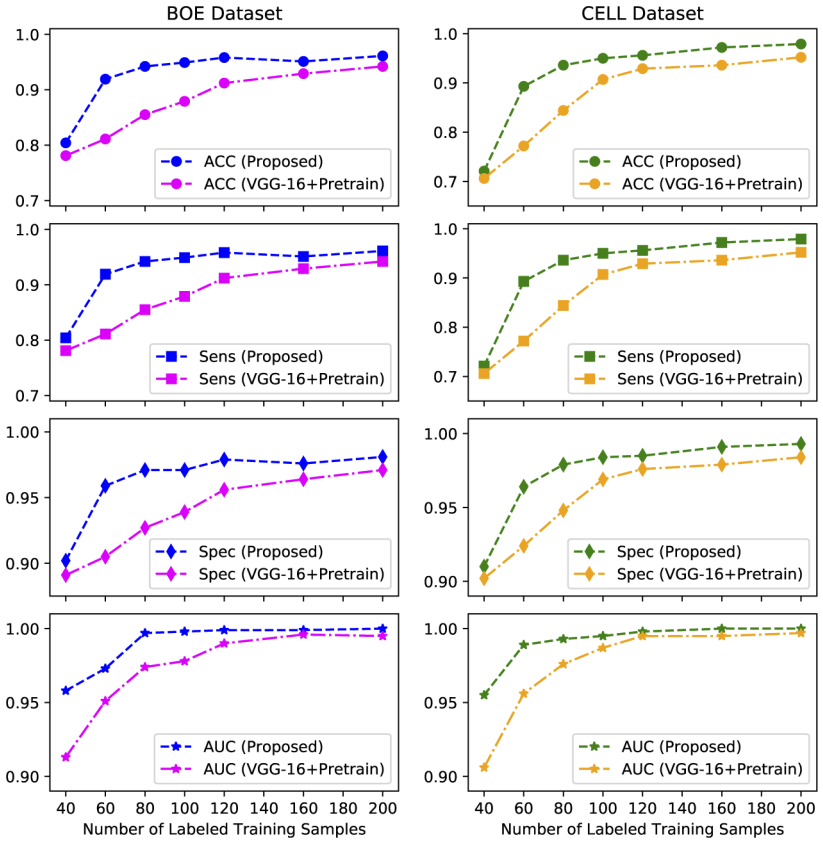


Fig. 8. The performance of the proposed method with different numbers of labeled OCT images. "ACC", "Sens" and "Spec" stand for accuracy, sensitivity and specificity, respectively.

Biomedical Optics EXPRESS

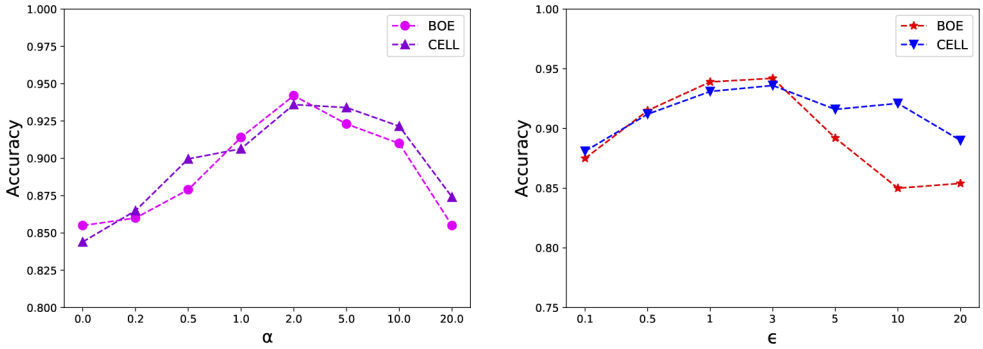


Fig. 9. The performance of the proposed method with different α and ϵ values on the two datasets.

found that the proposed method with only 80 labeled OCT images can achieve expert level when identifying retinopathy (the lowest recognition accuracy of experts is 0.921.). With 200 labeled OCT images, the proposed method outperforms four out of six human experts, which indicates the effectiveness of the proposed method. Besides, it also proves that the proposed automatic retinopathy detection system can be used for clinical applications.

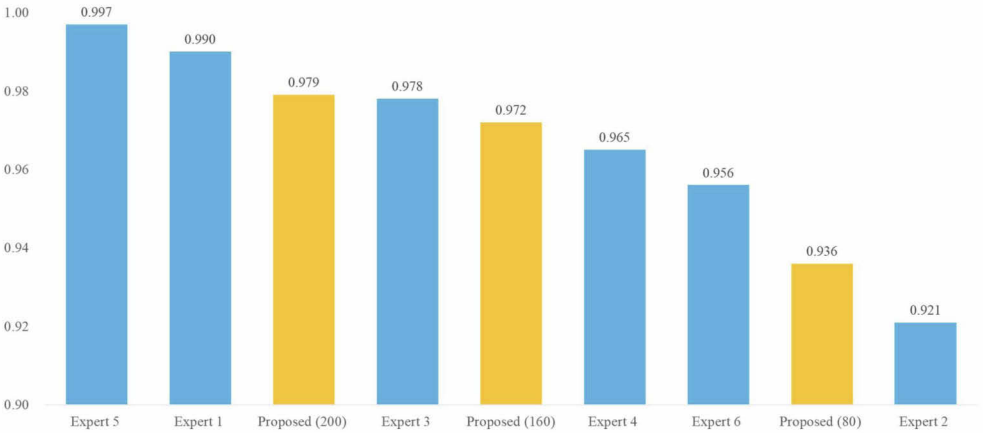


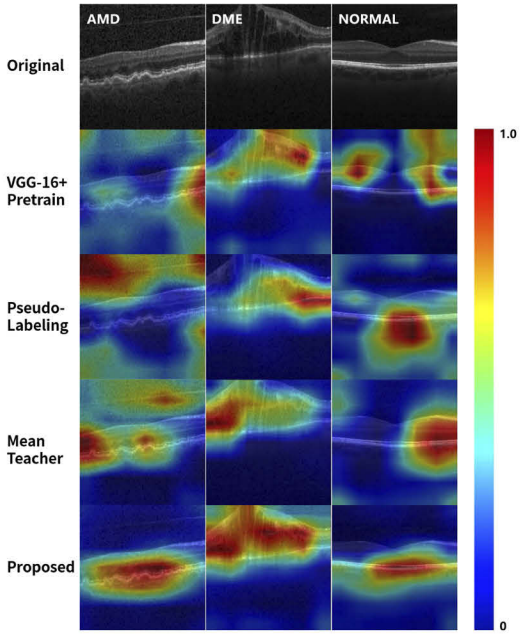
Fig. 10. Comparison between the proposed method with 80, 160 and 200 labeled OCT images for training and human experts.

3.7. Visualization with Grad-CAM

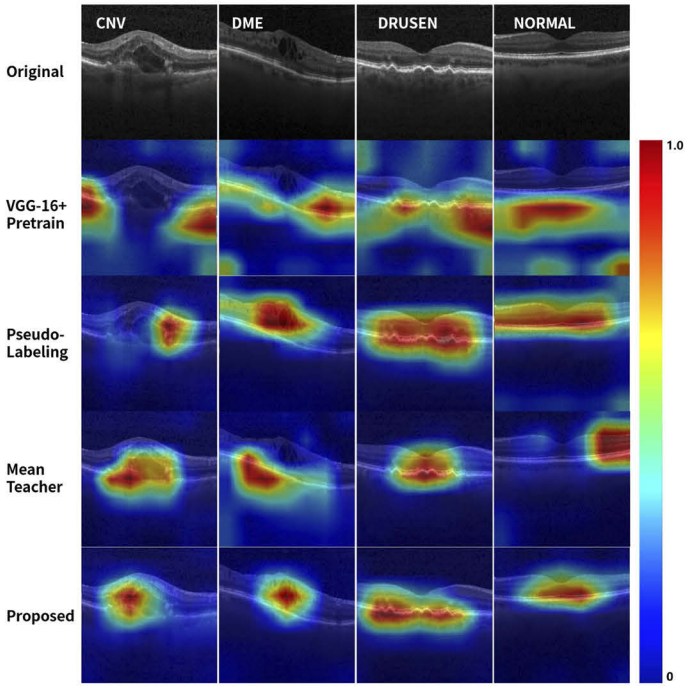
In order to interpret the predictions of the models, we adopt the Grad-CAM technique to show the saliency maps of the proposed method, the VGG-16 with pretraining, pseudo-labeling and mean teacher (Pseudo-labeling and mean teacher are the second best models in Table 2 and Table 1 respectively). Some representative samples are manually selected, illustrated in Fig. 11. It can be found that the VGG-16 with pretraining often finds the wrong regions of the OCT images when making predictions, leading to a poor performance. Generally, semi-supervised learning methods have a superior performance on finding the key patterns. Among them, the proposed method performs the best. It can accurately find the key patterns of the OCT images when predicting different retinopathies. This explains why the proposed method performs the best for retinopathy

Biomedical Optics EXPRESS

detection. Besides, it will also give confidence to ophthalmologists when using the proposed system for the detection of retinopathy with OCT images.



(a) BOE dataset



(b) CELL dataset

Fig. 11. The visualization of VGG-16 with pretraining, pseudo-labeling, mean teacher and the proposed method by using Grad-CAM.

Biomedical Optics EXPRESS

We need to emphasize that there are also some bad cases where the models cannot find the key patterns. These cases may lead to wrong predictions. We have manually selected two bad cases (true label is DME) in the two datasets, shown in Fig. 12. The network prediction for each method has been indicated on the figure. It can be found that all the four models cannot well capture the key patterns. In this case, the models will have a much higher chance to give wrong predictions. For example, the mean teacher focuses on a normal area of the sample from BOE, resulting a wrong prediction as NORAML. Similarly, the VGG-16 with pretraining and pseudo-labeling also predict DME as NORMAL when focusing on a normal area of the OCT image from CELL. The proposed method and mean teacher predict the DME as DRUSEN as they focus on a region that looks like drusen for the given sample from CELL.

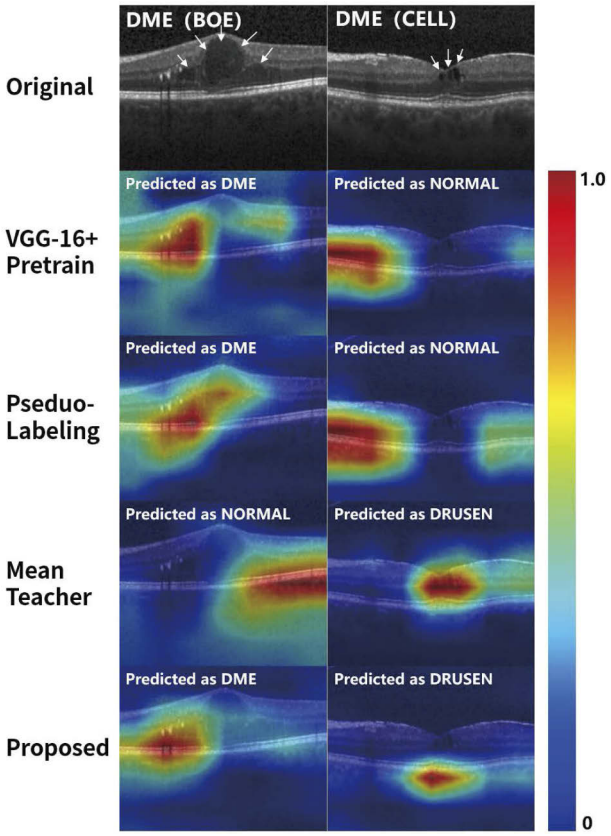


Fig. 12. Bad examples for the two datasets. White arrows indicate the key patterns.

4. Discussion

In this paper, we proposed a semi-supervised deep learning method for automatic detecting retinopathy with OCT images. The proposed method achieves recognition accuracies of 0.942 and 0.936 with only 80 labeled OCT images for training on two widely used datasets. Even though the benchmark approaches in the literature have shown great performance for the detection of retinopathies based on OCT images, they highly rely on the huge number of labeled OCT images for model training, which may not be available in real world scenarios. The proposed semi-supervised deep learning methods only requires very few labeled and large number of

Biomedical Optics EXPRESS

unlabeled OCT images, to achieve a high detection accuracy. We also verified that if more labeled OCT images are available, the performance of the proposed method can be further enhanced.

When comparing the proposed method with human experts, we surprisingly find that the proposed method with the training on only 80 labeled OCT images can achieve expert level in recognizing retinopathies with OCT images. By further increasing the number of labeled OCT images to 200, it can outperform four experts from six. Note that these human experts are with significant clinical experience in an academic ophthalmology center. This unique property of the proposed method is quite meaningful and practical, especially for developing countries where the number of ophthalmologists is inadequate for the diagnosis of retinopathy with OCT images.

Only giving an accurate prediction is not enough for clinical applications. If we are also able to provide some evidences on how the algorithm makes certain predictions, ophthalmologists will be confident on the prediction results. In this paper, we adopt the Grad-CAM approach to visualize the key regions (patterns) of the input OCT images that determine the final predictions. In experiments, it can be found that the proposed method is able to focus on the correct patterns when predicting retinopathies.

Overall, the proposed semi-supervised deep learning method for retinopathy detection only requires a very small amount of labelled OCT images which can be easily obtained. It is able to achieve expert level performance with only 80 labelled OCT images. Besides, it can provide some evidences for the prediction by finding key patterns of the OCT images. However, there are some limitations of the proposed method when applying to real-world clinical applications. First of all, since only very few labelled OCT images are adopted, class imbalance issue (e.g., normal class may be dominate in real applications) may have a big influence on model performance. Another key issue is the quality of these few labeled OCT images. If the quality of OCT images cannot be guaranteed, the performance of the proposed method would suffer a lot.

In our future works, we will consider to further enhance the performance of the proposed method by using optical coherence tomography angiography (OCTA) images which contain additional information for retinopathy detection [31,32].

5. Conclusion

In this paper, we proposed a semi-supervised deep learning method for retinopathy detection with OCT images. The proposed method consists of a pretrained VGG-16 network for feature learning on raw OCT images and a virtual adversarial training (VAT) to incorporate large amount of unlabeled OCT images for performance improvement. In experiments, the proposed method outperforms benchmark approaches and achieves accuracies of 0.942 and 0.936 with only 80 labeled OCT images on two popular datasets. When comparing with human experts, the proposed method with 200 labeled OCT images performs better than four experts from six. By adopting the Gradient Class Activation Map (Grad-CAM) technique, we also illustrate the saliency maps which indicate the key regions of the input OCT images when making predictions. The proposed method can accurately find the key patterns in input OCT images, which verifies its superior performance on retinopathy detection.

Funding. National Medical Research Council Individual Research Grant (MOH-OFIRG19may-0009); Ministry of Education - Singapore Academic Research Fund Tier 1 (2018-T1-001-144); Ministry of Education - Singapore Academic Research Funding Tier 2 (MOE-T2EP30120-0001).

Disclosures. The authors declare no conflicts of interest related to this article.

References

1. M. Treder, J. L. Lauermann, and N. Eter, "Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning," Graefe's Arch. Clin. Exp. Ophthalmol. **256**(2), 259–265 (2018).
2. X. Li, L. Shen, M. Shen, F. Tan, and C. S. Qiu, "Deep learning based early stage diabetic retinopathy detection using optical coherence tomography," Neurocomputing **369**, 134–144 (2019).

Biomedical Optics EXPRESS

3. E. A. Swanson and J. G. Fujimoto, "The ecosystem that powered the translation of oct from fundamental research to clinical and commercial impact," Biomed. Opt. Express **8**(3), 1638–1664 (2017).
4. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**(7553), 436–444 (2015).
5. C. S. Lee, A. J. Tyring, N. P. Deruyter, Y. Wu, A. Rokem, and A. Y. Lee, "Deep-learning based, automated segmentation of macular edema in optical coherence tomography," Biomed. Opt. Express **8**(7), 3440–3448 (2017).
6. F. G. Venhuizen, B. van Ginneken, B. Liefers, F. van Asten, V. Schreur, S. Fauser, C. Hoyng, T. Theelen, and C. I. Sánchez, "Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography," Biomed. Opt. Express **9**(4), 1545–1569 (2018).
7. A. G. Roy, S. Conjeti, S. P. K. Karri, D. Sheet, A. Katouzian, C. Wachinger, and N. Navab, "Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," Biomed. Opt. Express **8**(8), 3627–3642 (2017).
8. H. Fu, M. Baskaran, Y. Xu, S. Lin, D. W. K. Wong, J. Liu, T. A. Tun, M. Mahesh, S. A. Perera, and T. Aung, "A deep learning system for automated angle-closure detection in anterior segment optical coherence tomography images," Am. J. Ophthalmol. **203**, 37–45 (2019).
9. C. S. Lee, D. M. Baughman, and A. Y. Lee, "Deep learning is effective for classifying normal versus age-related macular degeneration oct images," Ophthalmol. Retin. **1**(4), 322–327 (2017).
10. D. Wang and L. Wang, "On oct image classification via deep learning," IEEE Photonics J. **11**(5), 1–14 (2019).
11. R. Rasti, H. Rabbani, A. Mehridehnavi, and F. Hajizadeh, "Macular oct classification using a multi-scale convolutional neural network ensemble," IEEE Trans. Med. Imaging **37**(4), 1024–1034 (2018).
12. F. Li, H. Chen, Z. Liu, X.-d. Zhang, M.-s. Jiang, Z.-z. Wu, and K.-q. Zhou, "Deep learning-based automated detection of retinal diseases using optical coherence tomography images," Biomed. Opt. Express **10**(12), 6204–6226 (2019).
13. D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, and F. Yan, "Identifying medical diagnoses and treatable diseases by image-based deep learning," Cell **172**(5), 1122–1131.e9 (2018).
14. J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, and D. Visentin, "Clinically applicable deep learning for diagnosis and referral in retinal disease," Nat. Med. **24**(9), 1342–1350 (2018).
15. S. P. K. Karri, D. Chakraborty, and J. Chatterjee, "Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration," Biomed. Opt. Express **8**(2), 579–592 (2017).
16. F. Li, H. Chen, Z. Liu, X. Zhang, and Z. Wu, "Fully automated detection of retinal disorders by image-based deep learning," Graefe's Arch. Clin. Exp. Ophthalmol. **257**(3), 495–505 (2019).
17. G. Ding, S. Zhang, S. Khan, Z. Tang, J. Zhang, and F. Porikli, "Feature affinity-based pseudo labeling for semi-supervised person re-identification," IEEE Transactions on Multimed. **21**(11), 2891–2902 (2019).
18. Y. Yao, J. Deng, X. Chen, C. Gong, J. Wu, and J. Yang, "Deep discriminative cnn with temporal ensembling for ambiguously-labeled image classification," in *AAAI*, (2020), pp. 12669–12676.
19. A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, (2017), pp. 1195–1204.
20. T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," IEEE Trans. Pattern Anal. Mach. Intell. **41**(8), 1979–1993 (2019).
21. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, (2017), pp. 618–626.
22. P. P. Srinivasan, L. A. Kim, P. S. Mettu, S. W. Cousins, G. M. Comer, J. A. Izatt, and S. Farsiu, "Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images," Biomed. Opt. Express **5**(10), 3568–3577 (2014).
23. D. Kermany, K. Zhang, and M. Goldbaum, "Large dataset of labeled optical coherence tomography (oct) and chest x-ray images," Mendeley Data, v3 http://dx.doi.org/10.17632/rscbjbr9sj **3** (2018).
24. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition* (Ieee, 2009), pp. 248–255.
25. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556 (2014).
26. J. Wang, G. Deng, W. Li, Y. Chen, F. Gao, H. Liu, Y. He, and G. Shi, "Deep learning for quality assessment of retinal oct images," Biomed. Opt. Express **10**(12), 6057–6072 (2019).
27. O. Perdomo, S. Otálora, F. A. González, F. Meriaudeau, and H. Müller, "Oct-net: A convolutional network for automatic classification of normal and diabetic macular edema using sd-oct volumes," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, (IEEE, 2018), pp. 1423–1426.
28. T. Shanthi and R. Sabeenian, "Modified alexnet architecture for classification of diabetic retinopathy images," Comput. & Electr. Eng. **76**, 56–64 (2019).
29. A. V. Varadarajan, P. Bavishi, P. Ruamviboonsuk, P. Chotcomwongse, S. Venugopalan, A. Narayanaswamy, J. Cuadros, K. Kanai, G. Bresnick, M. Tadarati, S. Sipa-archa, J. Limwattanayingyong, V. Nganthavee, J. R. Ledsam, P. A. Keane, G. S. Corrado, L. Peng, and D. R. Webster, "Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning," Nat. Commun. **11**(1), 130 (2020).

Biomedical Optics EXPRESS

30. S. Ruuska, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen, and J. Mononen, "Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle," Behav. Processes **148**, 56–62 (2018).
31. E. Vaghefi, S. Hill, H. M. Kersten, and D. Squirrell, "Multimodal retinal image analysis via deep learning for the diagnosis of intermediate dry age-related macular degeneration: a feasibility study," J. Ophthalmol. **2020**, 1–7 (2020).
32. P. Zang, L. Gao, T. T. Hormel, J. Wang, Q. You, T. S. Hwang, and Y. Jia, "Dcardnet: Diabetic retinopathy classification at multiple levels based on structural and angiographic optical coherence tomography," IEEE Transactions on Biomed. Eng. (2020).